# A Bioinformatics Method Identifies Prominent Off-targeted Transcripts in RNAi Screens

Published in final edited form as:
*Nat Methods*. ; 9(4): 363–366. doi:10.1038/nmeth.1898.

# A Bioinformatics Method Identifies Prominent Off-targeted Transcripts in RNAi Screens

**Frederic D. Sigoillot**[1], **Susan Lyman**[1,2], **Jeremy F. Huckins**[1,3], **Britt Adamson**[4], **Eunah Chung**[1], **Brian Quattrochi**[1,5], and **Randall W. King**[1,*]

[1]Department of Cell Biology, Harvard Medical School, Boston, Massachusetts, United States of America

[4]Department of Genetics, Harvard Medical School, Boston, Massachusetts, United States of America

## Abstract

Because off-target effects hamper interpretation and validation of RNAi screens, we developed a bioinformatics method, Genome-wide Enrichment of Seed Sequence matches (GESS), to identify candidate off-targeted transcripts from direct analysis of primary screening data. GESS identified a prominent off-targeted transcript in several screens, including *MAD2* in a screen for components of the spindle assembly checkpoint. We demonstrate how incorporation of the results of GESS analysis can enhance the validation rate in RNAi screens.

RNA interference (RNAi) is a powerful discovery tool, but frequent false positives complicate analysis of genome-wide RNAi screens[1–3]. The problem arises because siRNAs can induce microRNA-like effects, downregulating expression of hundreds of genes nonspecifically[4,5]. Strikingly, such effects can occur with as few as 6–7 nucleotides of sequence complementarity, although effects may become more pronounced with greater complementarity[6]. Some transcripts may be particularly susceptible to off-targeting[7–9], but the identification of such transcripts typically occurs only after much effort has been expended to validate genes of interest. Therefore, new methods are necessary to identify off-targeted transcripts earlier in the validation process.

We conducted an image-based high-throughput siRNA screen (Supplementary Results 1 and Supplementary Fig. 1) to identify novel components of the spindle assembly checkpoint (SAC)[10]. We determined that off-target effects were pervasive, as we were unable to validate any novel genes from the primary screen despite identifying known components of the pathway. To understand the basis of the off-target effect, we tested 34 siRNAs with the strongest phenotype for their ability to downregulate known components of the SAC, and

*Correspondence should be addressed to R.W.K. (randy_king@hms.harvard.edu).
[2]Current Address: Gilead Sciences, Inc., Foster City, California, United States of America
[3]Current Address: Department of Psychology and Brain Sciences, Dartmouth College, Hanover, New Hampshire, United States of America
[5]Current Address: Center for Academy Achievements, University of Massachusetts Medical School, Worcester, Massachusetts, United States of America

Note: Supplementary information is available on the Nature Methods website.

**AUTHOR CONTRIBUTIONS**

F.D.S., S.L. and R.W.K. conceived the study. F.D.S., S.L., E.C., B.Q. and B.A. performed the experiments. F.D.S. and J.F.H. wrote the GESS program code. F.D.S. and R.W.K. wrote the manuscript.

**COMPETING FINANCIAL INTERESTS**

The authors declare no competing financial interests.

found that all 34 siRNAs strongly decreased *MAD2* mRNA and protein levels in addition to their intended target (Supplementary Results 2 and Supplementary Fig. 2). Half of these siRNAs contained a 7mer seed sequence complementary to the *MAD2* 3′UTR, indicating the potential for microRNA-like off-targeting. We tested seven of these seed-match containing siRNAs, and found that all could downregulate a *MAD2* 3′UTR reporter construct (Supplementary Fig. 3). We found that over half of all 324 active siRNAs in the screen contained a 7mer seed match in the *MAD2* 3′UTR sequence, whereas only 8% of the inactive siRNAs contained a seed match. These findings indicate that the majority of active siRNAs in our SAC screen are likely to produce a phenotype by nonspecifically targeting the *MAD2* transcript.

To identify such potentially devastating off-target effects prior to the validation process, we developed an approach that utilizes primary screening data to identify transcripts that are sensitive to off-target effects (Fig. 1). Phenotypic screen data is used to separate the siRNAs into two groups: “with phenotype” and “without phenotype”. The program then calculates the seed match frequency (SMF) for active ($SMF^a$) and inactive ($SMF^i$) siRNAs for each transcript encoded in the genome (Fig. 2). In principle, transcripts that are sensitive to off-targeting will bias the ratio of $SMF^a$: $SMF^i$ (Seed Match Enrichment, or SME) such that it exceeds one and the statistical significance of this bias relative to other genes in the data set is determined. We refer to this approach as Genome-wide Enrichment for Seed Sequence match (GESS) analysis. It can be performed using genome-wide databases of full-length mRNAs or sub-regions of mRNAs (3′UTRs, 5′UTRs, coding sequence), although we have only identified off-targeted genes using the 3′UTR database, consistent with known rules of microRNA-based targeting.

We first evaluated the ability of GESS to identify *MAD2* as an off-targeted transcript in our spindle checkpoint screen. We applied GESS analysis to compare the seed match frequency of the most active siRNAs that produced a loss of SAC phenotype ($n = 49$) to the siRNAs that did not ($n = 9{,}856$). We analyzed each of 27,534 3′UTR sequences in the human genome (Fig. 2a). When using a 7mer seed match from either the antisense or sense strand seed sequences of an siRNA as a search criterion, we found that the 3′UTR of the *MAD2* transcript showed a significant seed match enrichment ($SMF^a$: $SMF^i$) of 8 fold ($SMF^a$ = 65.3%; $SMF^i$ = 8.2%; $P = 4.2\times10^{-23}$). The only other significantly enriched transcript represented another *MAD2* sequence in the database. A GESS analysis where all siRNA seed sequences were randomly scrambled showed no statistically significant outliers (Supplementary Fig. 4).

We determined how the GESS analysis of our SAC screen was affected by the following set of parameters: i) strength of phenotype; ii) the seed sequence length, iii) the seed match multiplicity; iv) the source of inactive control siRNAs; and v) seed sequence strand choice (Supplementary Results 3). Relaxing the phenotype strength led to identification of additional outliers, yet *MAD2* remained the most statistically enriched transcript (Supplementary Fig. 5). Increasing the stringency of the method by lengthening the seed from 7 to 8 nucleotides also permitted specific identification of *MAD2* (Supplementary Fig. 6). Increasing the seed match multiplicity, which increases stringency by requiring two seed matches per transcript, failed to identify *MAD2* in some cases (Supplementary Fig. 7). Because most published RNAi screens do not provide the nucleotide sequences of all tested siRNAs, we developed an alternative method for generating a set of inactive seed sequences, in which nucleotide 1 of the seed sequences of active siRNAs was changed to its complement (P1c-seeds), and found that this method could be used as a source of inactive siRNAs (Supplementary Fig. 8). Finally, considering seed matches from only the siRNA antisense strand showed better sensitivity but somewhat lower specificity than including each strand in the analysis (Supplementary Fig. 9).

We next tested whether GESS can identify off-targeted transcripts in other published screens. A recent screen identified siRNAs that could overcome mitotic arrest induced by a small-molecule inhibitor of the mitotic kinesin Eg5[11]. Since mitotic arrest induced by this mechanism is SAC-dependent[12], we anticipated that *MAD2* could be an off-target in this screen. In this case, the set of experimentally-determined inactive siRNAs was not published, so we used P1c-seeds to generate the set of inactive siRNAs. GESS analysis of this data set, using 7mer seeds and a seed match multiplicity of one, indeed identified the *MAD2* 3′UTR as the strongest statistically significant outlier, with an SME value of 3.9 ($P = 3.3\times10^{-18}$; Fig. 2b). A control analysis where all active and inactive siRNA seed sequences were randomly scrambled showed no significant outliers (Supplementary Fig. 10).

We tested GESS further on a previously published RNAi screen of 6,000 human genes for novel components of the TGFβ pathway, which failed to identify any novel components of the pathway and was plagued by off-target effects[9]. In that study, the vast majority of active siRNAs tested (89%; 172 of 193 tested) were experimentally confirmed to reduce mRNA levels of either the TGFβ Receptor 1 or 2, with the latter being more sensitive. We performed GESS analysis on the primary data of the screen, using the 391 active siRNAs and 18,869 inactive siRNAs. Using at least one 7mer seed match as a search criterion, GESS identified the *TGFβ-R2* transcript (represented by two sequences in the database) as the major outlier in the analysis with SME values of 1.6 ($P = 1.9\times10^{-12}$) and 1.4 ($P = 3.9\times10^{-9}$) while the *TGFβ-R1* transcript (two sequences in the database) showed no significant enrichment (SME = 0.97, $P = 0.664$ and SME = 0.99, $P = 0.832$) (Fig. 2c). A third weak outlier was identified but there is no evidence that it is involved in the TGFβ pathway. A control GESS analysis with randomly scrambled seed sequences for all siRNAs showed no significant outliers (Supplementary Fig. 11). We also investigated the effect of varying GESS parameters on the outcome of the analysis (Supplementary Results 4 and Supplementary Fig. 12).

Finally, GESS also identified RAD51 as a potential off-targeted gene in a screen for genes required for homologous recombination[13] and off-targeting of RAD51 was confirmed experimentally[13]. To examine whether GESS can help prioritize hits from siRNA screens, we investigated the consequences of removing siRNAs that contain a seed match against the RAD51 3′UTR (Fig. 3). The primary screen, followed by pool deconvolution, identified 88 candidate genes using a criterion of at least two of four siRNAs producing the phenotype. After removing siRNAs that contain a 7mer seed match to *RAD51* 3′UTR, 63 candidate genes retained at least two of four active siRNAs. We compared the performance of the original 88 candidates to the set of 63 “GESS-selected” candidate genes. Three additional independent siRNAs targeting these 88 genes were tested for their ability to reduce homologous recombination. In this analysis, 32 of 88 genes scored with at least two of three additional siRNAs, a confirmation rate of 36%. When the analysis was restricted to the set of GESS-selected candidates, 32 of 63 candidate genes were positive (51%). None of the 25 candidates eliminated by GESS showed a phenotype with more than one out of three new siRNAs. When this process was repeated using ten randomly selected genes containing a 3’UTR of similar length, no positive effect on validation rate was observed (Supplementary Table 1). This analysis indicates the value of taking into account potential off-target transcripts identified by GESS in prioritizing genes for validation in siRNA screens.

There is no tool other than GESS, to our knowledge, that can systematically examine screening data to directly identify potential off-targeted transcripts. A previously described approach to identify off-target effects in screens searches for siRNA seed sequences that are statistically overrepresented in the set of active siRNAs as compared to inactive siRNAs[7,8,14], but does not identify which transcripts might be targeted. We compared GESS

to seed sequence enrichment analysis. For screens in which GESS identified a biologically confirmed, statistically significant outlier, we attempted to identify 7mer seeds that were overrepresented in active siRNAs compared to inactive siRNAs (Supplementary Table 2a). In our SAC screen, we identified 8 such seed sequences (Supplementary Table 2b), indicative of a potential off-target effect. However, seed sequence enrichment analysis alone failed to highlight the extent of off-targeting in the screen, as only 35 out of 324 active siRNAs (11%) contained a seed sequence that was significantly enriched. Furthermore, this analysis cannot directly identify the MAD2 3′UTR as the relevant off-target in this dataset. Analysis of the data set from the TGFβ pathway RNAi screen[9] identified one 7mer seed sequence that was significantly enriched among active siRNAs (Supplementary Table 2c), present in only five of 391 (1.28%) active siRNAs as compared to five of 18,869 inactive siRNAs (0.03%). Importantly, analysis of the Eg5 inhibitor override screen[11], as well as the homologous recombination screen[13], failed to identify statistically overrepresented seed sequences. In summary, GESS appears to be more sensitive in identifying potential off-target effects compared to simple seed sequence analysis, and is furthermore capable of directly identifying the sensitive transcript(s). Because GESS does not require that active siRNAs contain a common seed sequence, it can detect off-target effects even if no particular seed sequence is enriched among active siRNAs. GESS uses the sequence of an mRNA transcript to “integrate” the information that is contained among different active siRNAs.

In total, we have analyzed thirteen different screens (Supplementary Table 3), and identified four screens, described here, in which statistically significant outliers were identified, and for which microRNA-based off-targeting has been established as problematic. Nine published RNAi screen datasets showed either no significant outliers or a few weakly significant outliers whose biological significance has not been investigated. The sequences of inactive siRNAs were not published for five of these nine screens, and thus we relied on use of P1c-seed sequences as a source of inactive siRNAs. However, this approach is not as information rich as using the experimentally determined inactive siRNAs, because the statistical significance of enrichment in the GESS analysis depends not only on an increase in the frequency of seed matches to a transcript among active siRNAs, but also a corresponding decrease in frequency of seed matches among inactive siRNAs. Furthermore, GESS analysis of genome-wide screens is most informative if siRNAs are screened individually rather than as pools. Because screens in *Drosophila* and *C. elegans* utilize multiple siRNAs generated from long dsRNAs (~500 base pairs), GESS is unlikely to be informative in these systems.

Why some transcripts are particularly sensitive to miRNA-like off-targeting remains unclear. *MAD2* is average among known spindle checkpoint in terms of 3′UTR length or AU-richness, ruling out trivial explanations. The *MAD2* 3′UTR may contain specific secondary structures or bind to specific proteins that render it particularly sensitive to off-target effects. Alternatively, the SAC may be particularly sensitive to small changes in *MAD2* protein levels. Consistent with this idea, *MAD2* is a haplo-insufficient tumor suppressor in vivo, and cells lacking one copy of *MAD2* show decreased ability to arrest in mitosis in the presence of microtubule inhibitors[15]. Similarly, the process of homologous recombination may be particularly sensitive to RAD51 gene dosage, explaining why RAD51 was identified by GESS as a prominent off-target in an siRNA screen for genes involved in homologous recombination[13]. Finally, similar observations were reported for the TGFβ pathway RNAi screen[9] where minor reductions of the TGFβ receptor transcripts appear to have major effect on the screen assay. Together these findings suggest it may be useful to assemble a database of genes whose transcripts are highly sensitive to off-target effects, and incorporate this information into the design algorithms used to generate siRNAs. Incorporation of GESS as a routine component of the analysis of high-throughput screens

should enable investigators to counter-screen for downregulation of sensitive transcripts and reduce the false positive rate during the validation process. Identification of transcripts sensitive to off targeting will also enable a better understanding of the rules that govern miRNA-like targeting and help further improve the design of siRNA reagents for future RNAi screens.

## METHODS

The GESS standalone package used in this manuscript is provided as a compressed archive (Supplementary Software 1–5). Software packages for updated versions will be available on our website (http://king.med.harvard.edu/). All siRNA sequences and associated phenotype data used to perform GESS analyses described in this manuscript are provided as a compressed archive (Supplementary Data 1). Excel result files for the main GESS analyses in this manuscript are provided as a compressed archive (Supplementary Data 2). Transcript database files for the human and mouse genomes are available as a compressed archive (Supplementary Data 3). Methods and associated references are available in the online version of the paper at http://www.nature.com/naturemethods/.

## ONLINE METHODS

### Tissue culture

HeLa H2B-GFP cells were grown from low passage in DMEM (Cell-Gro, Cat#10-013-CV) supplemented with 10% Fetal Bovine Serum (FBS; Atlanta Biologicals Cat#S11150), in a humidified incubator at 37°C and 5% $CO_2$.

### siRNA library

The Qiagen “Druggable” genome siRNA library V1.0, consisting of two individual siRNAs for 5,090 human genes, was used in our primary siRNA screen. Non-targeting siRNA control #3 (D-001210-01-20) and *MAD2* (GGAACAACUGAAAGAUUGG-dTdT, custom synthesis), were from Dharmacon. The sequences of all siRNAs used in this study are reported in supplementary excel files along with corresponding phenotypic data.

### siRNA transfections and image-based screening

HeLa H2B-GFP cells were plated at 1,000 cells per well in 30 μl OptiMEM containing 1% FBS, supplemented with penicillin, streptomycin and 2 mM glutamine (pen/strep/glu) in 384-well plates (Corning, #3712), 16–18 hrs prior to transfection. The cells were washed twice with OptiMEM containing pen/strep/glu, and then incubated in 40 μl OptiMEM pen/strep/glu per well for 1–4 hrs prior to transfection. For each well to be transfected with siRNA, 8 μl OptiMEM, 0.5 μl GTS diluent and 0.25 μl GeneSilencer (Genlantis) was first premixed, and then added to 2 μM siRNA. The siRNA-reagent mix was incubated at room temperature for 15 min and then added to cells, yielding a final siRNA concentration of 150 nM. Four-to-six hours post-transfection, 20 μl DMEM containing 30% FBS were added. Taxol (150 nM final concentration) was added to the cells 32–36 hrs post-transfection. Twenty-four hours later, cells were fixed and nuclei stained by adding one volume of DPBS fix/stain solution (final concentrations: 3.7% Formaldehyde, 250 ng/ml Hoechst 33342 (Molecular Probes H-3570) and 0.1% Triton X-100). After 20 min incubation at room temperature, the cells were washed 2–3 times with DPBS. Fluorescence images of nuclei were obtained using a CellWoRx high-content screening microscope (Applied Biosystems). Nuclear morphology was analyzed by manual inspection of images. Untransfected cells and those transfected with control siRNA remain arrested in mitosis under these conditions (Supplementary Fig. 1). In contrast, cells treated with a positive control siRNA targeting the essential SAC component MAD2 exited mitosis, as indicated by the presence of interphase

cells with multilobed nuclei (SAC bypass; Supplementary Fig. 1). Each siRNA was transfected in duplicate wells and each well was imaged in one location. Each image was given a penetrance phenotype (P) reflecting the number of cells affected by the siRNA. The penetrance categories were: 3 (80–100% cells affected), 2.5 (60–80% cells), 2 (40–60% cells), 1.5 (15–40% cells) and 1 (>0% to 15% cells). A sub-rating (SR), reflecting the proportion of affected cells showing SAC bypass, was also assigned using similar categories from 3 to 1. The penetrance and sub-rating category values were multiplied to reflect the overall rate of bypass in each image and the higher rate of the two replicates per siRNA was retained. Three phenotype thresholds were considered in the present analyses: a high threshold ($P \times SR = 9$), yielding 49 active siRNAs; a relaxed threshold ($P \times SR \geq 7.5$), yielding 137 siRNAs; and a low threshold ($P \times SR \geq 2$), yielding 324 active siRNAs.

## Plasmid constructs

Total RNA was isolated and purified from HeLa H2B-GFP cells using the RNeasy kit (Qiagen Cat# 74104). A cDNA library was generated by reverse transcribing the total RNAs using a reverse transcription system (Promega, A3500) following manufacturer's protocol. *MAD2* mRNA sequences were PCR amplified from the cDNA library. The PCR primers contained XbaI sites at both extremities. XbaI digested PCR fragments were cloned into the pGL3-control vector (Clontech) digested with XbaI. This results in expression of an mRNA coding for the Firefly luciferase with *MAD2* sequences downstream of the stop codon. The BglII-BamHI cassette from pRL-TK vector (Clontech), containing the Renilla Luciferase gene under the control of HSV Thymidine kinase promoter, was non-directionally cloned into the BamHI site of the pGL3-control and pGL3-control-*-*MAD2* sequences vectors. Resulting plasmids sequences were verified by DNA sequencing.

## Luciferase reporter assays

HeLa H2B-GFP cells were plated in 24-well plates (BD Falcon 353047) at 30,000 cells per well in 500 µl OptiMEM containing 1% FBS and pen/strep/glu. Sixteen hours after plating, cells were transfected with 50nM siRNAs with GeneSilencer as follows. The cells were washed with OptiMEM and then incubated 1–4 hrs in 150 µl OptiMEM + pen/strep/glu in the absence of FBS. GTS diluent (2.5 µl) and 1.25 µl GeneSilencer reagent were premixed in 40 µl OptiMEM and added to 5 µl of siRNA stocks (2 µM) for each well. The siRNA transfection mix was incubated 15 min at room temperature and added to the cells. DMEM containing 20% FBS (200 µl) was added to each well 4–6 hrs after siRNA transfection. Twenty-four hours later, the siRNA transfection medium was replaced with 500 µl growth medium (without pen/strep/glu) and plasmid transfection was initiated. Plasmids (500 ng per well) were transfected with Fugene 6 (Roche) using a reagent ratio of 5 µl Fugene 6: 2 µg plasmid. OptiMEM (100 µl) was mixed with 0.75 µl Fugene 6 and pre-incubated at room temperature for 5 min. The pre-mix was added to 500 ng plasmid. The plasmid-reagent mix was then added to the cells after 15 min incubation at room temperature. Dual luciferase assays were performed 24–48 hrs after initiating plasmid transfections, following the manufacturer's protocol (Dual-Glo system, Promega). Luminescence measurements were performed on an Envision plate reader (Perkin Elmer).

## Branched DNA (bDNA) assay for mRNA level quantification

Messenger RNA levels were measured, in duplicate, 48 hrs after siRNA transfection of 20,000 HeLa H2B-GFP cells per well in 24-well plates (as described for the luciferase assays). The bDNA assay (QuantiGene / Panomics) was conducted following the manufacturer's protocol and using probe sets specific to *MAD2* (PA-11305-02; NM_002358), *BUBR1* (PA-11159-01; NM_001211), *BUB1* (PA-11577-01; NM_004336), or the housekeeping genes *GAPDH* (PA-10382-02; NM_002046) and *PPIB*/*Cyclophilin* (PA-10384-02; NM_000942). Duplicate measurements were averaged, normalized per

average housekeeping *PPIB* mRNA measurement. The normalized ratio for control siRNA transfected cells was used as 100% reference for determination of relative changes in the ratio for other siRNAs. The results were displayed as a heat map with indicated scale using Spotfire DecisionSite.

### Quantitative Western blotting

MAD2 and GAPDH protein levels were determined by SDS-PAGE separation of proteins followed by Western blotting. The proteins were detected using a rabbit anti-MAD2 (Bethyl, A300-301A) and mouse anti-GAPDH (AbCam, ab8245) antibodies. Secondary antibodies coupled to fluorophores (anti-mouse Alexa-Fluor750 and anti-rabbit Alexa-Fluor680, Invitrogen) were used to detect both signal on the same membrane using an Odyssey (Li-Cor Biosciences) scanner. Quantifications were performed using the Odyssey program and are reported as MAD2/GAPDH signal ratio, normalized to control treatment.

### Sequence databases

Genome-wide sequences for human 5′UTRs, coding sequences or 3′UTRs were retrieved from the Ensembl database using the online tool Martview (www.biomart.org; Ensembl Genes 61, Human genome built GRCh37 or earlier) or Refseq using the UCSC Genome Bioinformatics table browser (http://genome.ucsc.edu/cgi-bin/hgTables; Refseq 44). Sequences of 19 nucleotides or less and duplicate sequences were removed and the remaining sequences were formatted into a text file with one sequence per line and corresponding identity information was formatted into an excel file with the same name (Supplementary Data 3).

### Genome-wide Enrichment of Seed Sequences (GESS) bioinformatics tool

MATLAB 2007 and more recent versions were used to program and run the GESS seed match search tool. The program is provided as MATLAB m-code files and as standalone versions packaged with the appropriate MATLAB Compiler Runtime (MCR) for Windows 32 and 64bit, Linux 32 and 64bit and Mac OS (Supplementary Software 1–5). The packages were compiled using MATLAB Compiler version 4.14 or later. Input data files consist of two text files that can be generated following the GESS manual provided along with the program. One contains a list of either all 19 nucleotide siRNA sequences (sense strand sequence or target sequence) in upper case with ATGC code (no U) one sequence per line, or only the sequence of active siRNAs if inactive siRNA sequences are unavailable. The second file contains phenotypic data (1 for siRNA with phenotype, 0 for siRNA with no phenotype) in the same order as the siRNA sequences, one number per line when providing both active and inactive siRNA sequences. If only active siRNA sequences are provided, GESS generates control siRNA seed sequences by changing nucleotide 1 of each seed to its complement and no phenotypic data file is required. To run a GESS methodology negative control run, siRNA seed sequences of both active and inactive siRNAs can optionally be randomly scrambled by the program. The program requests the user to define the length of seed sequences to analyze (typically 6 to 8 nucleotides with the default being 7) and the minimal number of seed matches (multiplicity) an siRNA must show towards a target sequence in order to consider it matching. The program allows selection of the strand(s) of the siRNA that should be used in the analysis. Either strand is considered by default, meaning that a seed sequence derived from either the sense or antisense strand must contain a seed match to the target sequence for the siRNA to be considered matching. Alternatively, the antisense strand only, the sense strand only or both strands (each strand must satisfy the seed matching parameters) can be analyzed. The program also asks the user to indicate which genome-wide transcript sequence database text file should be used (3′UTR, 5′UTR and CDS sequence databases for human and mouse are provided as Supplementary Data 3). The transcript sequences must be in upper case, one sequence per line with ATGC code (no

U). Three different multiple hypotheses testing correction methods can be selected in the analysis, as below. If a significant outlier is detected in the primary GESS run, the analysis can be repeated after removing the major outlier, as this approach might enable other, less prominent off-target effects to be detected. The user simply chooses the option to exclude siRNAs matching a sequence and provides a text file containing the outlier sequence.

### Statistical analysis of GESS results

Because some of the sequences analyzed contained low seed match event numbers, we calculated the Chi square with correction for continuity (Yates' Chi square) statistics which compensates for low event numbers.

$$X^2_{Yates} = \frac{N \cdot (|N_{siPhenMatch} \cdot N_{siNoPhen} \cdot N_{siNoPhenMatch} \cdot N_{siPhen}| - N/2)^2}{(N_{siPhen} \cdot N_{siNoPhen} \cdot N_{siMatch} \cdot N_{siNoMatch})}$$

N: total number of siRNAs tested in the GESS analysis.

$N_{siPhen}$: number of siRNAs with phenotype.

$N_{siNoPhen}$: number of siRNAs with no phenotype.

$N_{siPhenMatch}$: number of siRNAs with phenotype with seed matching to tested sequence.

$N_{siNoPhenMatch}$: number of siRNAs with no phenotype with seed matching to tested sequence.

$N_{siMatch}$: number of siRNAs with seed matching to tested sequence.

$N_{siNoMatch}$: number of siRNAs with no seed matching to tested sequence.

$N_{siPhenNoMatch}$: number of siRNAs with phenotype with no seed matching to tested sequence.

$N_{siNoPhenNoMatch}$: number of siRNAs with no phenotype with no seed matching to tested sequence.

The one-tailed probability (*P* value) of the Yates' Chi square statistics was calculated (with a degree of freedom of one). The Chi Square was set to zero if the Chi Square calculation denominator was null (the Yates' Chi square cannot be calculated). The corresponding *P* value is then equal to one. If any of $N_{siPhenMatch}$, $N_{siPhenNoMatch}$, $N_{siNoPhenMatch}$, $N_{siNoPhenNoMatch}$ was less than or equal to 20, the Fisher's exact test two-sided *P* value was determined instead of the Yates's Chi Square *P* value. The genomic sequences were ranked from the one with lowest *P* value (rank = 1) to the one with highest *P* value (rank = A, the number of genomic sequences analyzed).

Three multiple hypotheses testing correction methods have been implemented in GESS. The Benjamini and Hochberg False Discovery Rate correction[16] (Simes' method) was used as default as it is considered a good balance between limiting report of false positive and false negative off-target transcripts. The null hypothesis (there is no difference between the frequency of siPhen and siNoPhen containing a seed match to a given sequence) was rejected if the *P* value calculated above was less than the corrected *P* value threshold ($\alpha \times$rank of sequence / A) where $\alpha$ is set as 0.05 by default (more stringent $\alpha$ values can be input by the user) and A is the number of genomic sequences analyzed. The number of

sequences passing or failing the test is indicated on each graph. Two additional methods are available for analysis by the user, namely, the Bonferroni[17] and the Bonferroni step-down[18] (Holm) methods. The corrected *P* value thresholds are (α / A) and (α / (A + 1 – rank of sequence)), respectively. These methods are more stringent than the Benjamini and Hochberg method. While they can be used to limit the rate of false positive off-targets identified, weaker genuine off-targets may be missed as false-negatives in the analysis. Corrected *P* value thresholds and associated statistical significance status for the three methods are reported in the *GESS_Results* file.

### Data visualization

The program plots the percentage of siRNAs containing a seed match to a transcript of interest, comparing the siRNAs with phenotype (Y-axis) to those without phenotype (X-axis). Each genomic sequence is represented by one point on the graph and statistical enrichment of significance is indicated in red. Alternatively, Spotfire DecisionSite was used to generate the graphs. Sequences with statistically significant seed match enrichment were depicted in red while non-significant sequences were depicted in gray. The numbers of significant and non-significant outliers are provided.

### siRNA seed sequence enrichment analysis (SSEA)

The GESS algorithm was adapted to be applied to siRNA seed sequences as follows. A list of 16,384 7mer seed sequences was generated and stored as a text file and excel file in the same format as the transcript sequences database files. These text files were used instead of the genome-wide transcript sequence databases to search for seed presence in the active and inactive siRNAs. All calculations and statistical decisions were performed similarly as for the GESS method. Provided fewer events are expected to be counted as compared to a GESS analysis, the multiple hypothesis testing error correction was restricted to the Benjamini and Hochberg method.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

We would like to thank the Institute of Chemistry and Cell Biology and Caroline Shamu for providing siRNA sequences for hits from the Eg5-inhibitor RNAi screen as well as the use of facility equipment for our screening experiments. We thank Sridaran Natesan and Paul August (Sanofi-Aventis) for helpful discussions in early stages of this work. We thank Steve Elledge (Harvard Medical School) for helpful discussions and for critical reading of the manuscript. We thank James Ware at the Harvard Catalyst Biostatistics consulting group for help in devising the statistical analysis workflow in the present manuscript: funding for statistical analysis was supported in part by Grant Number 1 UL1 RR025758-01, Harvard Clinical and Translational Science Center, from the National Center for Research Resources; the content is solely the responsibility of the authors and does not necessarily represent the official views of the National Center for Research Resources or the National Institutes of Health. This research was funded by a Sanofi-Aventis grant and National Institute of Health grant GM66492 to RWK.

## REFERENCES

1. Jackson AL, Linsley PS. Nat Rev Drug Discov. 2010; 9:57–67. [PubMed: 20043028]
2. Mohr S, Bakal C, Perrimon N. Annu Rev Biochem. 2010; 79:37–64. [PubMed: 20367032]
3. Sigoillot FD, King RW. ACS Chem Biol. 2011; 6:47–60. [PubMed: 21142076]
4. Jackson AL, et al. Nat Biotechnol. 2003; 21:635–637. [PubMed: 12754523]
5. Tschuch C, et al. BMC Mol Biol. 2008; 9:60. [PubMed: 18577207]
6. Bartel DP. Cell. 2009; 136:215–233. [PubMed: 19167326]
7. Lin X, et al. Oncogene. 2007; 26:3972–3979. [PubMed: 17173063]

8. Lin X, et al. Nucleic Acids Res. 2005; 33:4527–4535. [PubMed: 16091630]
9. Schultz N, et al. Silence. 2011; 2:3. [PubMed: 21401928]
10. Musacchio A, Salmon ED. Nat Rev Mol Cell Biol. 2007; 8:379–393. [PubMed: 17426725]
11. Tsui M, et al. PLoS One. 2009; 4:e7339. [PubMed: 19802393]
12. Kapoor TM, Mayer TU, Coughlin ML, Mitchison TJ. J Cell Biol. 2000; 150:975–988. [PubMed: 10973989]
13. Adamson B, Smogorzewska A, Sigoillot FD, King RW, Elledge SJ. Nat Cell Biol. 2012
14. Sudbery I, Enright AJ, Fraser AG, Dunham I. BMC Genomics. 2010; 11:175. [PubMed: 20230625]
15. Michel LS, et al. Nature. 2001; 409:355–359. [PubMed: 11201745]

## ONLINE METHODS REFERENCES

16. Benjamini Y, Hochberg Y. J Roy Statist Soc Ser B. 1995; 57:289–300.
17. Bonferroni C. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze. 1936; 8:3–62.
18. Holm S. Scandinavian Journal of Statistics. 1979; 6:65–70.

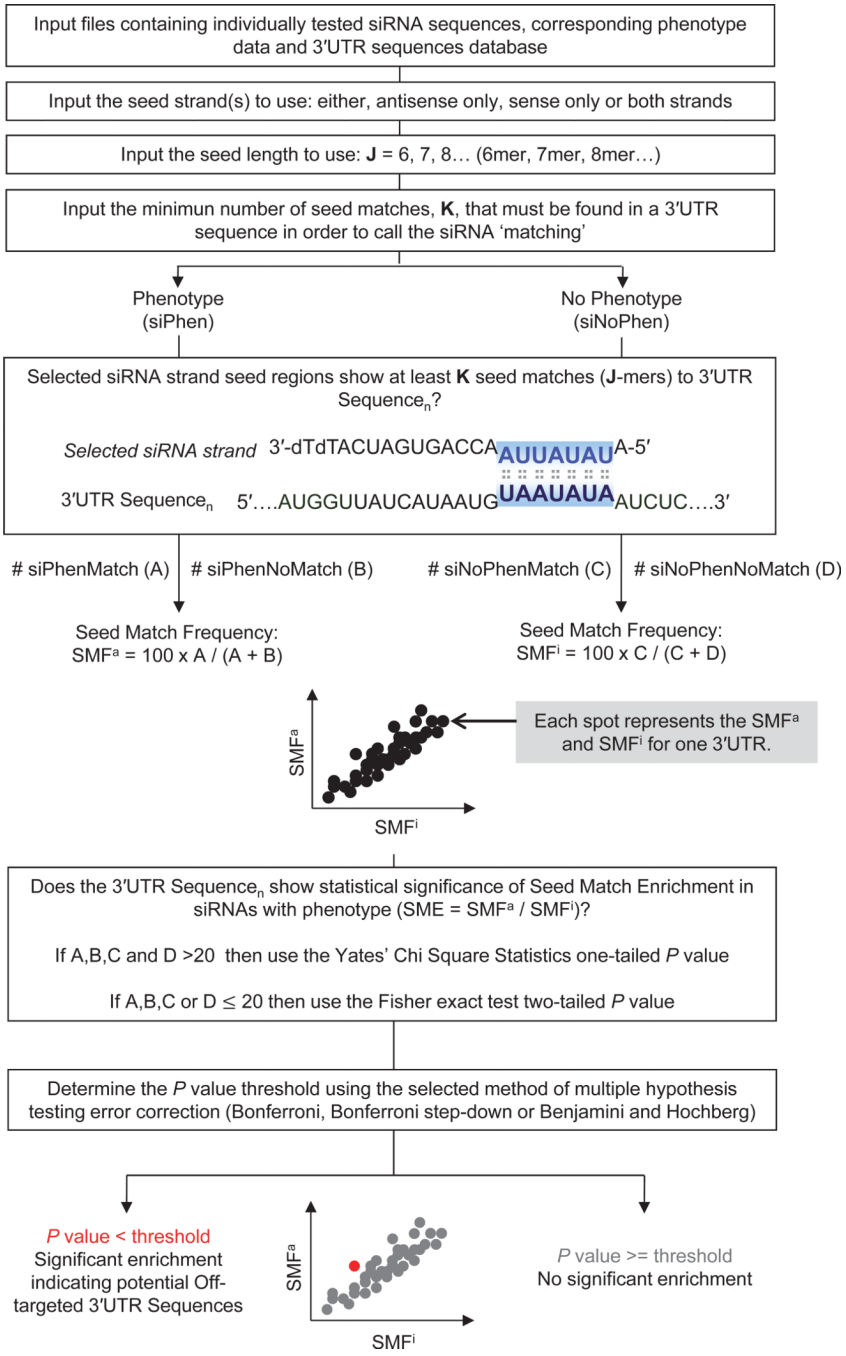


**Figure 1. Summary of the Genome-wide Enrichment for Seed Sequence matches (GESS) method**

The GESS algorithm begins by splitting the set of siRNAs into two sets: those with phenotype and those without phenotype. The user enters criteria for defining a matching transcript, including the siRNA strand(s), seed length (J) and seed match multiplicity (K). GESS calculates the percent of siRNAs in each set that shows seed matching with each sequence in the genome-wide database (SMF$^a$ for active siRNAs and SMF$^i$ for inactive siRNAs). Statistical significance of seed match enrichment (SME) among the set of active siRNAs compared to the set of inactive siRNAs is performed (see Methods).

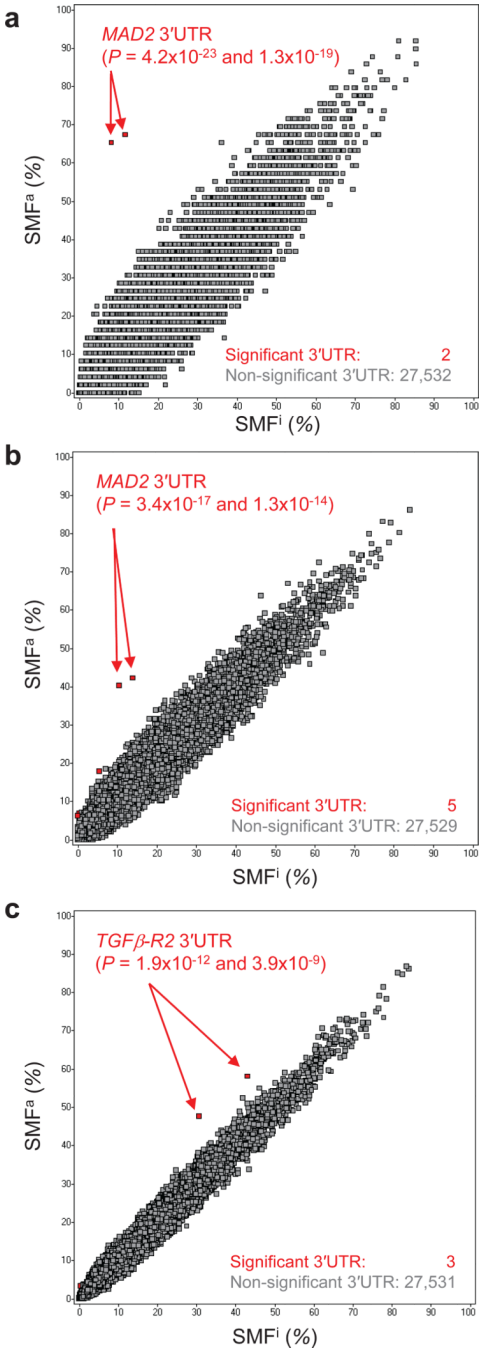


**Figure 2. GESS identifies major off-targeted transcripts in RNAi screen datasets**

(**a**) The plot shows GESS analysis of 27,534 human mRNA 3′UTRs on the primary data from a screen that identified siRNAs inducing loss of SAC function. Each point represents one 3' UTR, and indicates the percentage of active siRNAs containing a seed match to the 3' UTR ($SMF^a$; percent of $n = 49$ total active siRNAs) plotted against the percentage of inactive siRNAs containing a seed match to the 3' UTR ($SMF^i$; percent of $n = 9{,}856$ total inactive siRNAs). (**b**) The plot shows GESS analysis as above on data published from an siRNA screen for components required for mitotic arrest upon inhibition of the mitotic kinesin Eg5 in HeLa cells[11]. P1c-seeds were used as the source of inactive siRNAs ($n = 308$,

for both active and inactive siRNAs). (**c**) The plot shows GESS analysis as above, on data published from an siRNA screen for genes involved in the TGFβ pathway[9]. Experimentally identified siRNAs that showed no phenotype (a cutoff of two standard deviations of activity was used to separate active from inactive siRNAs) were used for the set of inactive siRNAs ($n$ = 391 active siRNAs, $n$ = 18,869 inactive siRNAs). Significance threshold was determined independently for each data point, using the Benjamini and Hochberg (Simes') method to correct the baseline value of α which is 0.05. Statistically significant outliers are depicted in red and their number is reported.

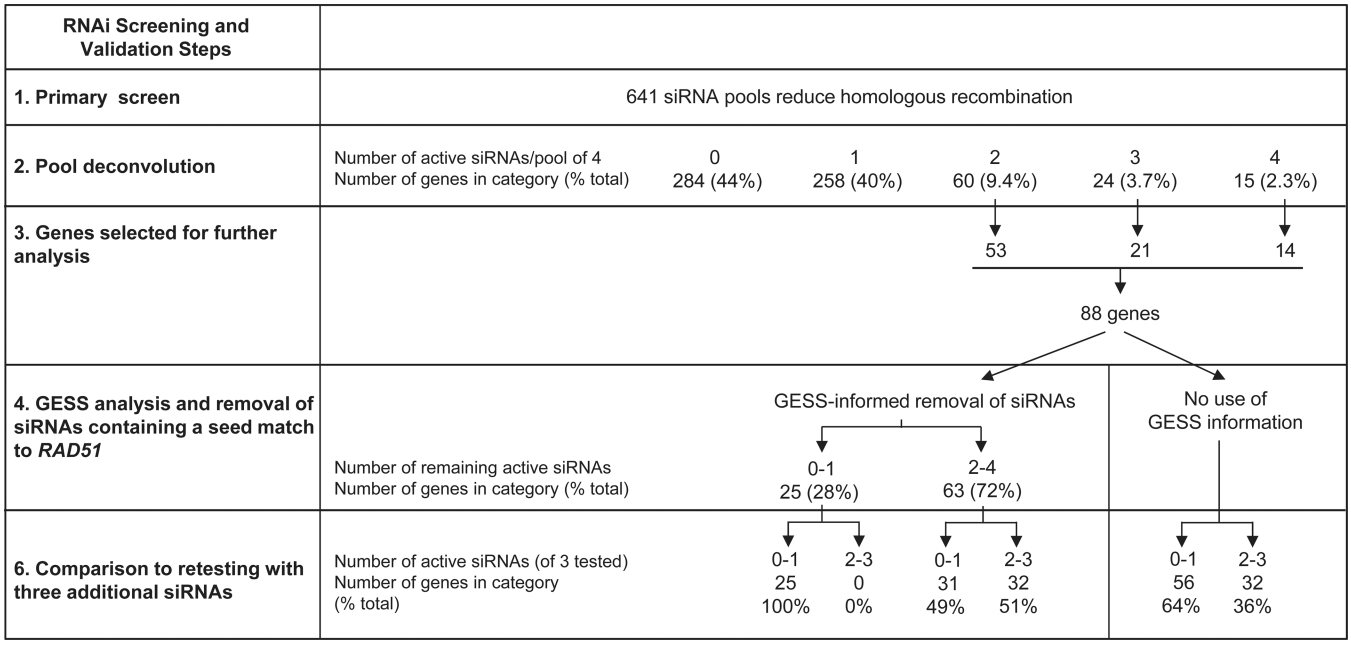


**Figure 3. GESS-informed selection of siRNA pools enriches for genes that reproduce the primary phenotype upon targeting with additional siRNAs**

The schematic shows that siRNA pools targeting 641 transcripts scored in a primary screen for genes required for homologous recombination. Upon deconvolution, pools targeting 99 genes showed the phenotype on at least two out of four siRNAs. Of these genes, 88 were further evaluated (11 genes were dropped because no additional siRNAs were commercially available, the original pool showed toxicity, or the retested genes were in the lower spectrum of primary screen scores). GESS analysis showed that the RAD51 3′UTR is sensitive to off-targeting. The schematic shows the rate at which the phenotype was reproduced with and without removal of siRNAs that contain a 7mer seed match to RAD51 3′UTR.